FORECASTING CONVENTIONAL GAS PRICES

ECON 4115 Final

Introduction

The goal of this project is to create an accurate forecasting model for the price of gasoline in the U.S. using its own time series data and data on crude oil prices. The forecasts for this data could be used by consumers to help make decisions for purchasing and using gasoline. These forecasts could also be used by economists and other researchers in tandem with other forecasting models to predict the status of the economy.

When reviewing other literature, it appears that most existing forecasts try to predict natural gas and crude oil prices, without a focus on consumer fuel prices. One paper used consumer survey from the Michigan Consumer Survey data to forecast gas prices, since survey data before has out-performed traditional time-series forecasting models. In this analysis, they found that this survey data has performed similarly to other forecasting methods, and did better in crisis periods, like the 2008 recession (Anderson et al., 2011). Another study incorporated oil prices into forecasting models for conventional gas prices to determine the value brought by the new econometric portion. Forecasts with the oil prices were compared to pure time-series forecasts, and they were found to not perform significantly better (Bastianin et al., 2014). Looking at studies on natural gas forecasting, one opted to use machine learning algorithms rather than more traditional forecasting methods. This study concluded that some ML models do outperform time-series models, but that these differences are small and require other algorithms to preselect variables (Čeperić et al., 2017).

Summary of the Data

The first time series tracks the average gas prices in the U.S. from 1991 to 2021. This data is collected weekly (every Monday at 8:00 am) across the nation and measured in dollars per gallon. Approximately 900 retail outlets are surveyed for self-serve gas prices, and these are averaged to provide the reported values. This data was obtained from the FRED database and is not seasonally adjusted.

The second time series, intended to be used in conjunction with the first, is of the price of crude oil from 1986 to 2021. This data is collected daily and is measured in dollars per barrel. West Texas Intermediate is used as the benchmark oil for the prices. This data was also obtained from FRED and is not seasonally adjusted.

There are 1598 observation in the dataset. Data was collected from January 28th, 1991, to September 6th, 2021. Forecasts generated from this data will be tested against new data up through December 6th, 2021, for conclusions.

Summary Statistics

Time Series	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation
Gas	0.885	1.192	2.026	2.051	2.713	4.054	0.862
Oil	-36.98	19.85	33.03	44.43	63.46	145.31	28.74

Introductory Analysis

The regular time series graph (Fig 1) indicates an upward trend and increasing volatility, but it is unclear if any seasonality is present. Seasonality was explored through a time series decomposition (Fig 2) and a seasonal-subseries plot with weekly frequency (Fig 3). These figures do indicate some higher prices in the summer months, however this is of a small magnitude, so it is still unclear if this should be built in. At this point, the volatility seen in the plot rules out simple forecasting methods (naïve, mean, trend) from consideration.

Next, the gas prices were compared to crude oil prices to determine the viability of oil prices as an explanatory variable for regression models. This was done by plotting the two variables against each other (Fig 4) and imposing the two series on top of each other (Fig 5) to find similar patterns. These plots indicate a strong relationship between these two variables. Some form of a dynamic regression model should be built for this relationship.

Models

Based on the introductory analysis, three models were built for forecasting gas prices; an exponential smoothing model, an ARIMA model, and a dynamic regression model with ARIMA errors. Each of these models used orders that were chosen by the statistical packages that created them, and changes were made if they were deemed necessary.

The first model was an ETS(M,Ad,N) model. The formal equation for this model is as shown below:

$$y_t = (l_{t-1} + \phi b_{t-1})(1 + \varepsilon_t)$$

$$l_t = (l_{t-1} + \phi b_{t-1})(1 + \alpha \varepsilon_t)$$

$$b_t = \phi b_{t-1} + \beta (l_{t-1} + \phi b_{t-1}) \varepsilon_t$$

I agreed with this form of the model. I did not think there was much seasonality present, and the errors seemed to get larger over time. The dampened trend may perform better than a standard one also because of the increasing volatility. The estimations for each parameter are as follows:

Parameter	Estimation
alpha	0.9435563
beta	0.8109462
phi	0.80
l_0	

	1.236779
b_0	-0.1067012

The model residuals (Fig 6) appear white noise and mean zero graphically, but the ACF of the residuals show significant values at lags 1 and 2.

The second model estimated was an ARIMA[(1,1,3)(0,0,0)] model with drift. The full equation in backshift notation is:

$$(1-\phi_1 B)(1-B)y_t = c + (1+\theta_1 B + \theta_2 B^2 + \theta_3 B^3)\varepsilon_t$$

The ACF and PACF plots of the differenced time series (Figs 7 & 8) were not easily decipherable, so the selection from the package was kept. However, a constant was added since a trend was visible in the data. The estimated parameters and their standard errors are:

Parameter	Estimation	Standard Error
phi_1	0.6248	0.0656
theta_1	-0.1638	0.0676
theta_2	-0.0230	0.0387
theta_3	0.0864	0.0321
c	5e-04	9e-04

Similar to the first model, the residuals plots (Fig 9) appear white noise. However, the significant lags in this case are at erratic lags and difficult to understand what could be causing this.

The third and final model used was a linear regression model with ARIMA(1,1,2) errors, as shown below:

$$y_t = \beta_1 oil + \eta_t$$

$$(1 - \phi_1 B)(1 - B)\eta_t = c + (1 + \theta_1 B + \theta_2 B^2)\varepsilon_t$$

The drift was removed from model 2 and replaced with the explanatory variable component. The estimated parameters were:

Parameter	Estimation	Standard Error
phi_1	0.7297	0.0400
theta_1	-0.2086	0.0480
theta_2	-0.0869	0.0405
oil	1e-04	3e-04

The residual plots for this model (Fig 10) showed significant correlations at the first 3 lags.

Model Selection

The best model was selected by calculating forecasts 12 weeks ahead with each model, and comparing these to the true values in the testing set of data. For the dynamic regression model, an ARIMA(0,1,2) model was built to forecast oil prices. These forecasts were used to calculate the dynamic regression model's forecasts. Fig 11 shows each forecast plotted with the true values laid over (CI intervals overlap). The table below shows the accuracy of each model based on numerous measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
ARIMA	0.0805	0.112	0.0955	2.45	2.94	0.288	0.238	0.822
dynreg	0.173	0.195	0.173	5.34	5.34	0.522	0.413	0.811
ETS	0.0376	0.0733	0.0666	1.12	2.06	0.201	0.155	0.824

Both the plot and the accuracy measures clearly indicate the exponential smoothing model as the best model. Fig 12 shows just this model compared to the actual values (with correct CI's). Just visually, this model does appear to do a great job following the data.

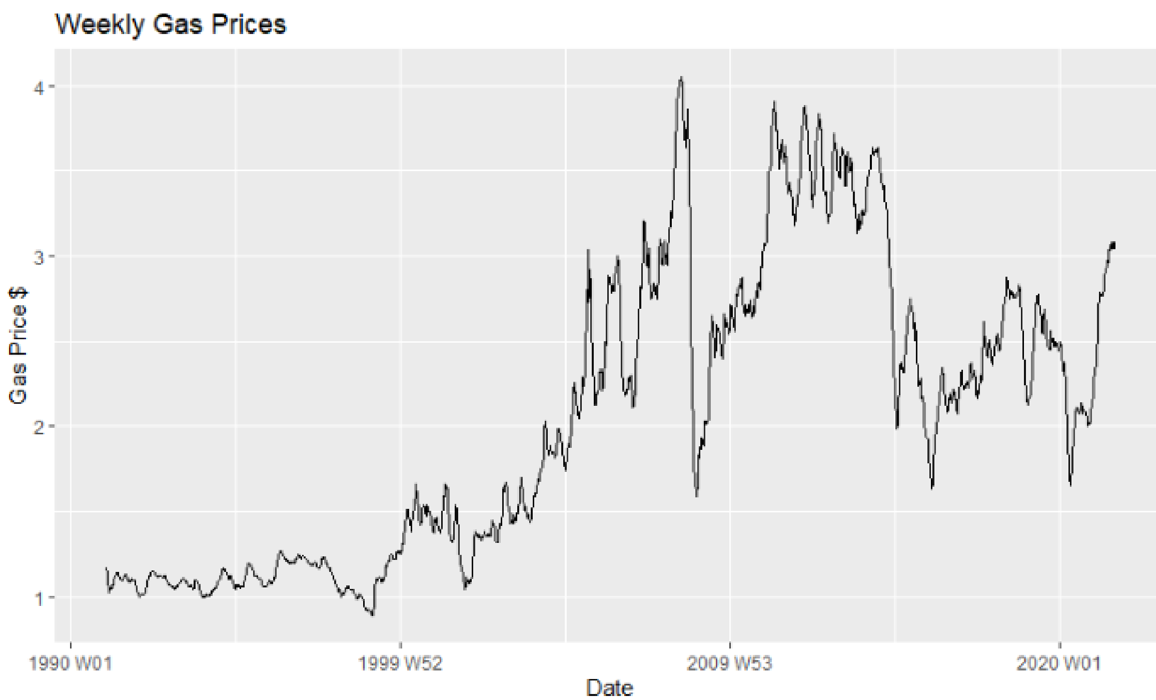
Conclusion

Based on knowledge learned in class, I was surprised to see the exponential smoothing model come out on top. ARIMA models seem to typically outperform ETS models, and the strong correlation between the gas and oil prices made me believe that the dynamic regression model would easily perform the best. I went back afterwards and used the actual oil values instead of the forecasted ones to see if the poor performance of the oil forecasts caused the model failure. However, the change in the accuracy was almost negligible. It may be that the greater simplicity of the exponential smoothing model was a strength in this case.

I believe that this model can be of use for following the trend of national gas prices, but will not do well in dealing with volatility as it comes up. Because of this, the forecast will be more useful for predicting macroeconomic impacts over longer periods, than as a way for consumers to predict prices in their own areas. Part of this has to do with the lack of seasonality seen in the national data, since I would have expected to see higher prices in the summer months (not just a difference of a couple of cents per gallon). In the future, it may be interesting to do this again, but with different sets of regional/local data to see how these differ from the national average.

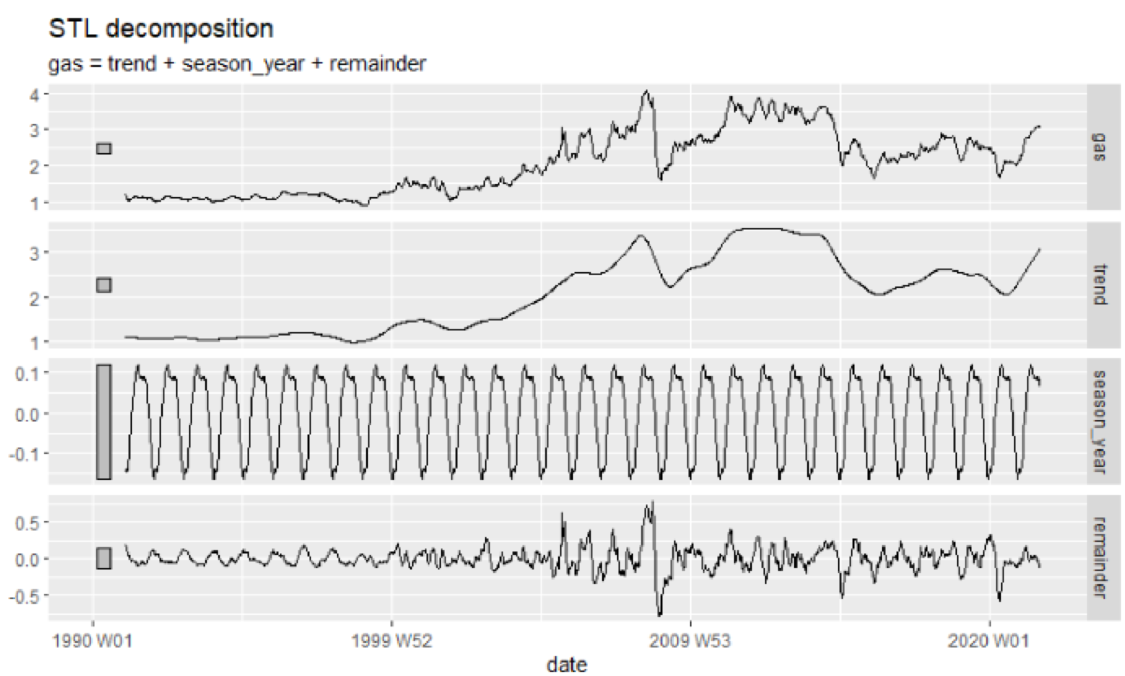
Appendix

Fig 1



Weekly Gas Prices
Gas Price $
4
3
2
1
1990 W01
1999 W52
2009 W53
2020 W01
Date

Fig 2



STL decomposition
gas = trend + season_year + remainder
gas
trend
season_year
remainder
0.1
0.0
-0.1
0.5
-0.5
1990 W01
1999 W52
2009 W53
2020 W01
date

Fig 3

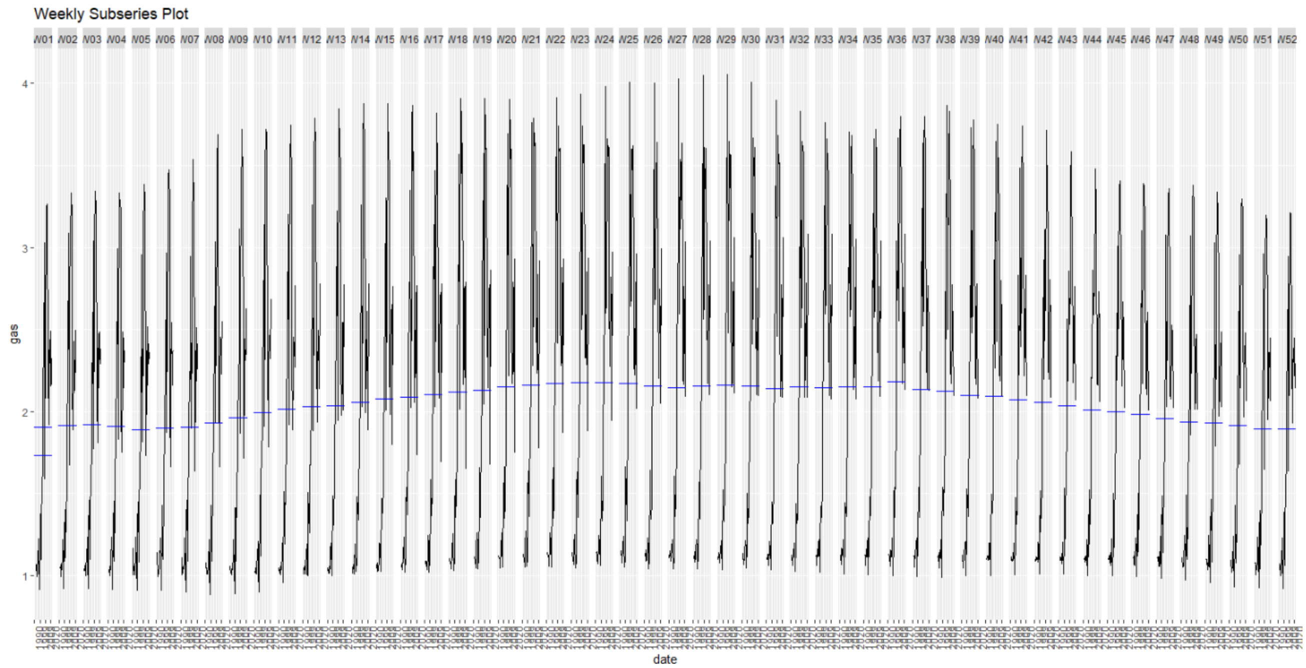

Weekly Subseries Plot
W01 W02 W03 W04 W05 W06 W07 W08 W09 W10 W11 W12 W13 W14 W15 W16 W17 W18 W19 W20 W21 W22 W23 W24 W25 W26 W27 W28 W29 W30 W31 W32 W33 W34 W35 W36 W37 W38 W39 W40 W41 W42 W43 W44 W45 W46 W47 W48 W49 W50 W51 W52
4
3
2
1
gas
date

Fig 4

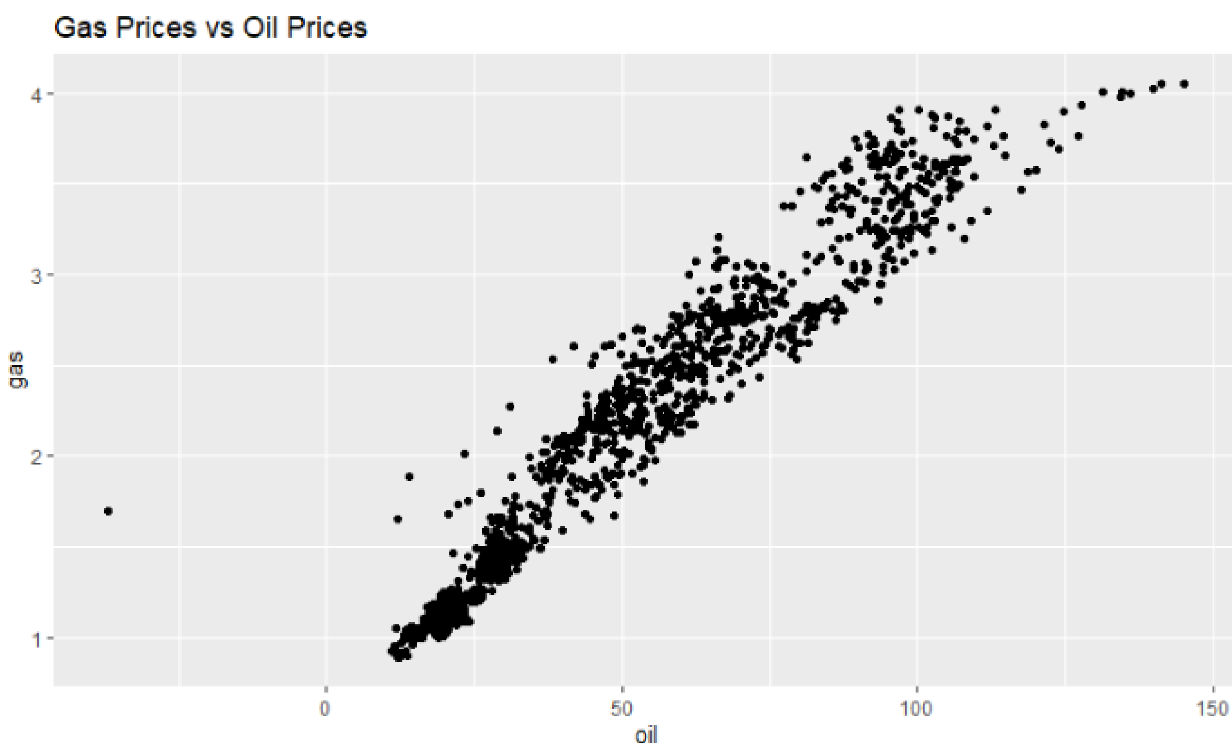

Gas Prices vs Oil Prices
4
3
2
1
gas
0
50
100
150
oil

Fig 5

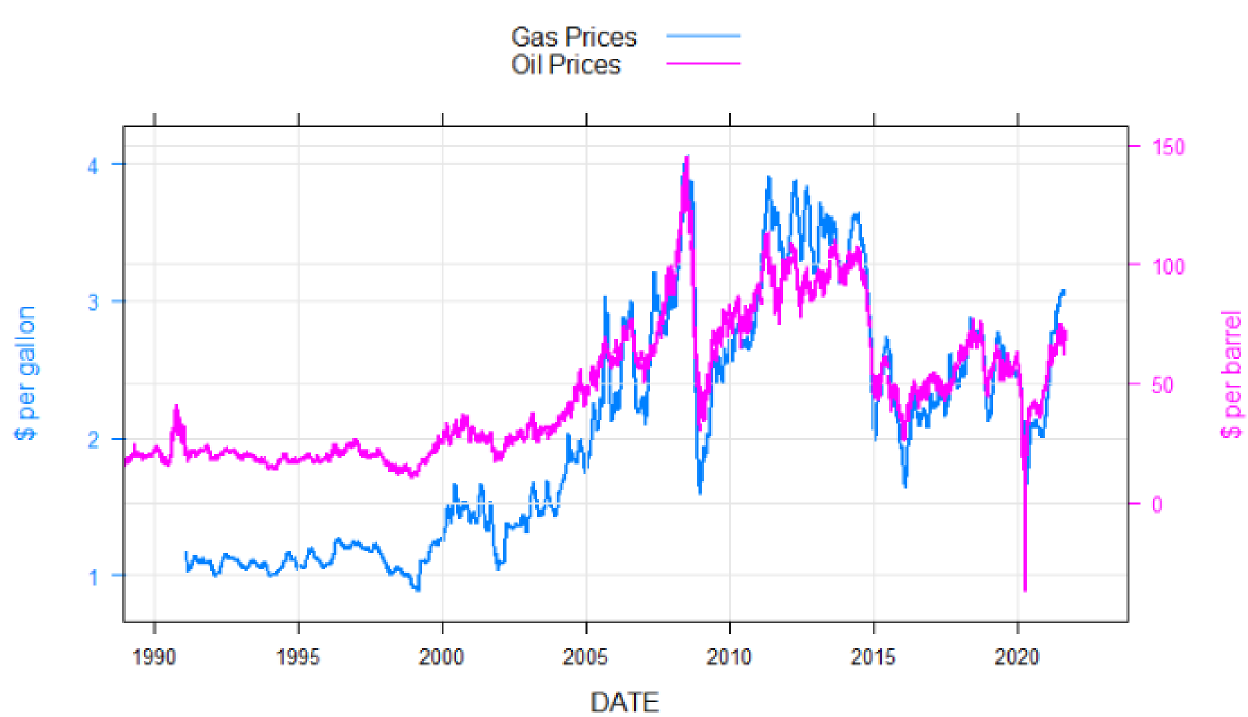

Gas Prices
Oil Prices
$ per gallon
$ per barrel
DATE
1990
1995
2000
2005
2010
2015
2020
4
3
2
1
150
100
50
0

Fig 6

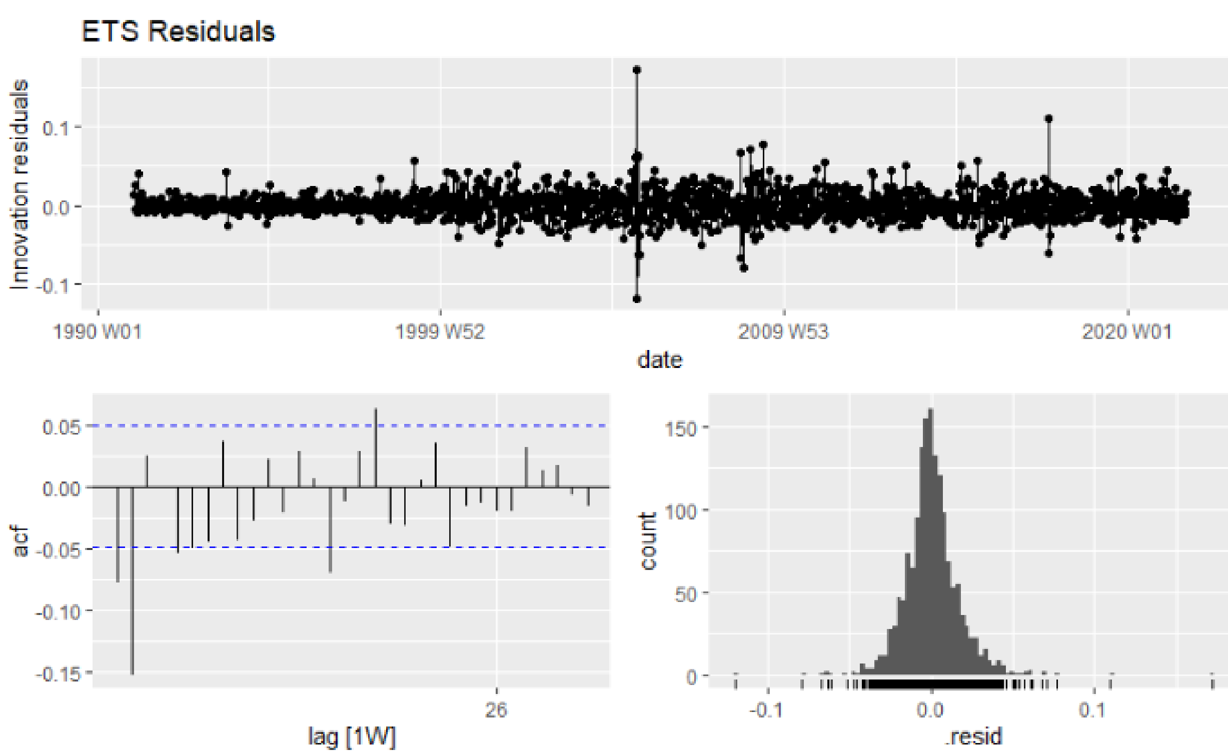
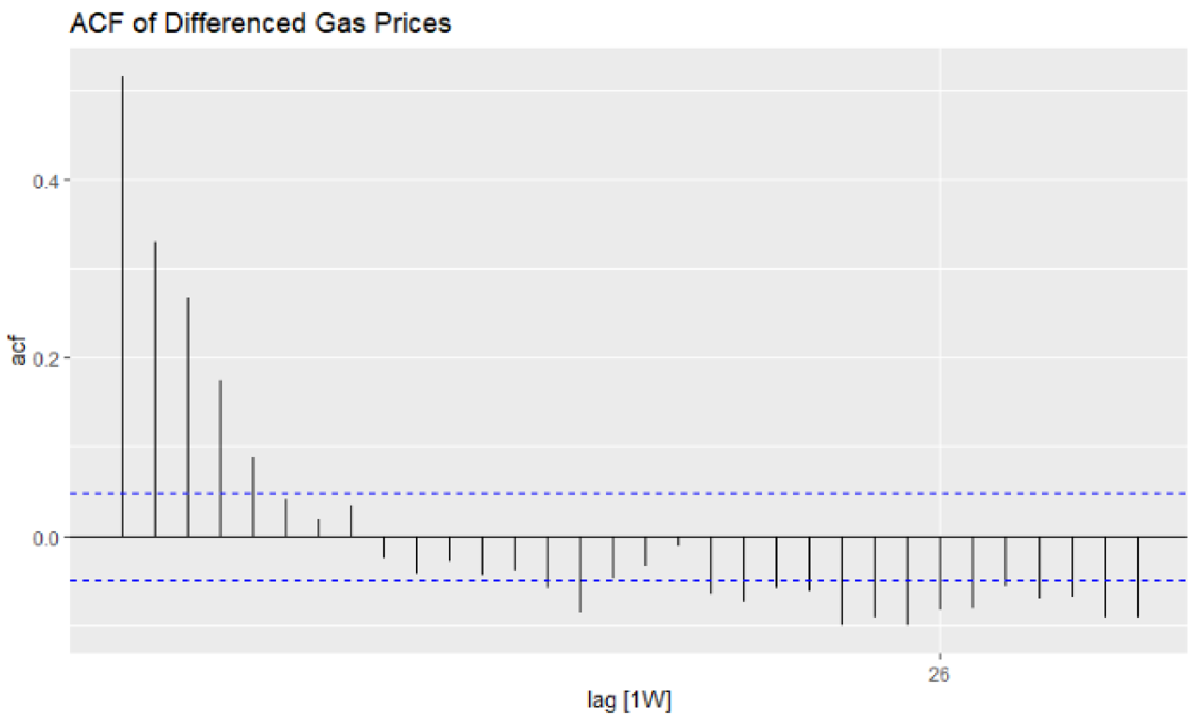
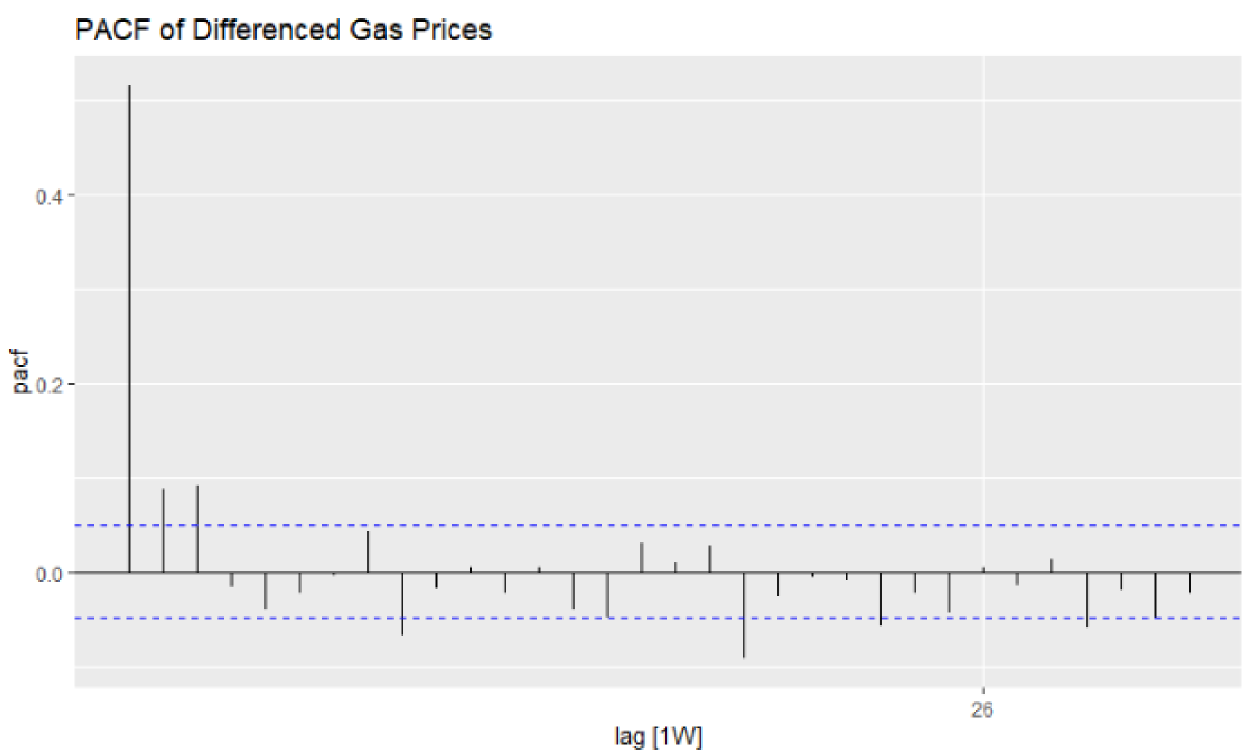

ETS Residuals
Innovation residuals
0.1
0.0
-0.1
1990 W01
1999 W52
2009 W53
2020 W01
date
acf
0.05
0.00
-0.05
-0.10
-0.15
26
lag [1W]
count
150
100
50
0
-0.1
0.0
0.1
.resid

Fig 7



ACF of Differenced Gas Prices
acf
0.4
0.2
0.0
26
lag [1W]

Fig 8



PACF of Differenced Gas Prices
pacf
0.4
0.2
0.0
26
lag [1W]

Fig 9

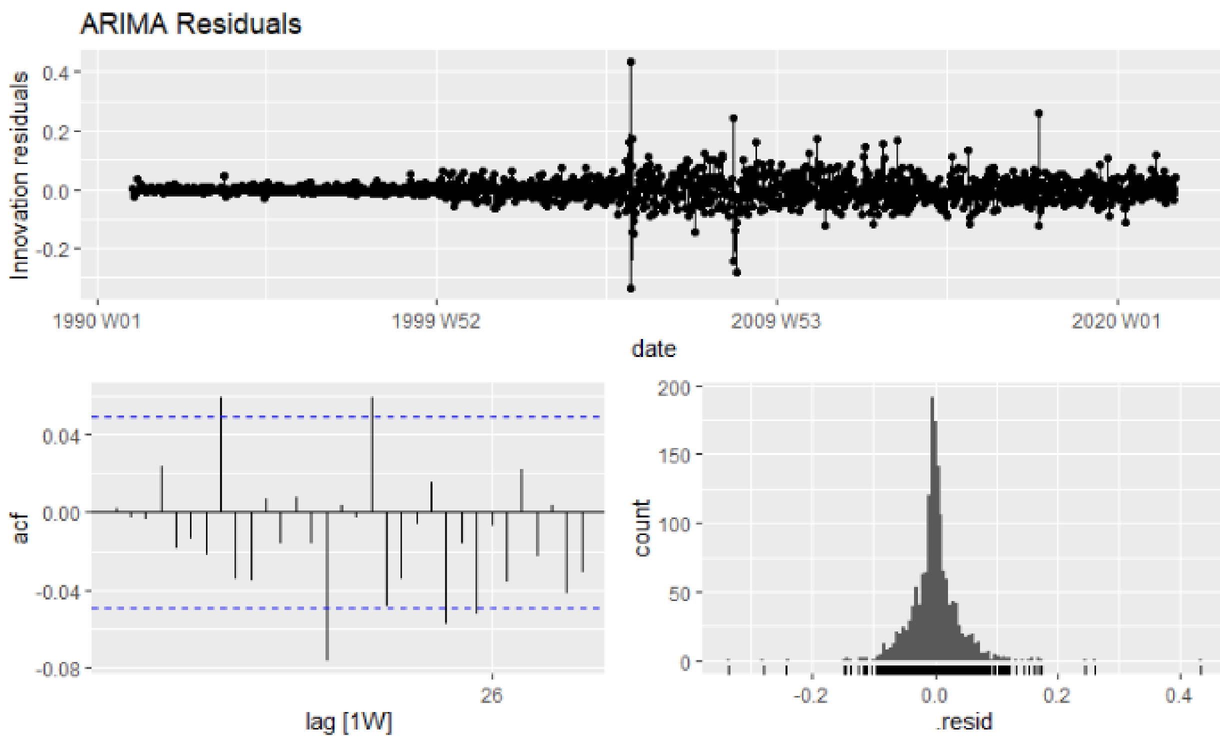

ARIMA Residuals
Innovation residuals
0.4
0.2
0.0
-0.2
1990 W01
1999 W52
2009 W53
2020 W01
date
acf
0.04
0.00
-0.04
-0.08
26
lag [1W]
count
200
150
100
50
0
-0.2
0.0
0.2
0.4
.resid

Fig 10

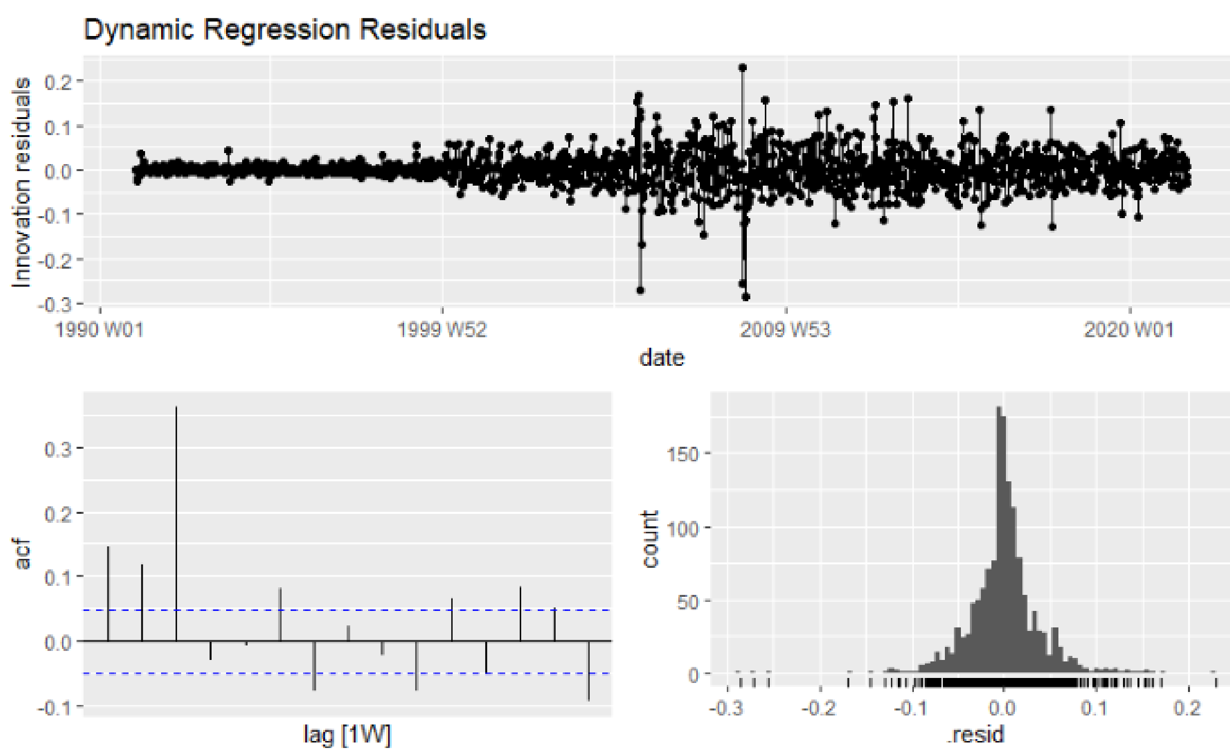

Dynamic Regression Residuals
Innovation residuals
0.2
0.1
0.0
-0.1
-0.2
-0.3
1990 W01
1999 W52
2009 W53
2020 W01
date
acf
0.3
0.2
0.1
0.0
-0.1
lag [1W]
count
150
100
50
0
-0.3
-0.2
-0.1
0.0
0.1
0.2
.resid

Fig 11

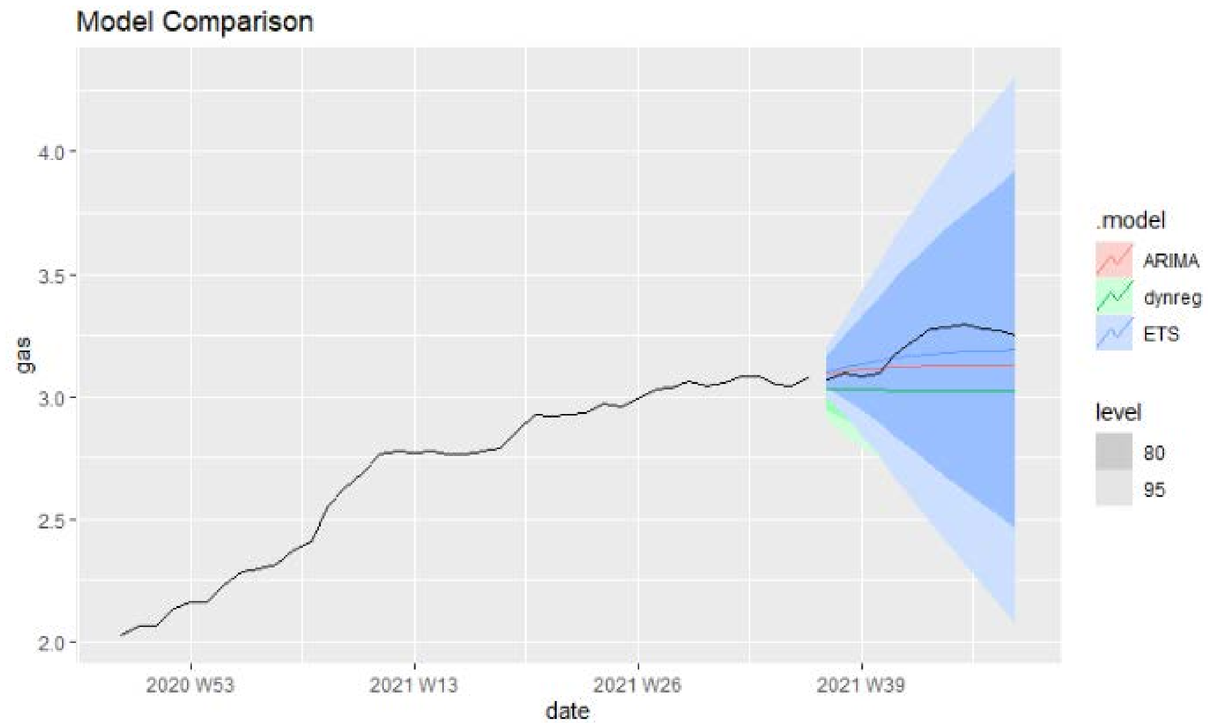
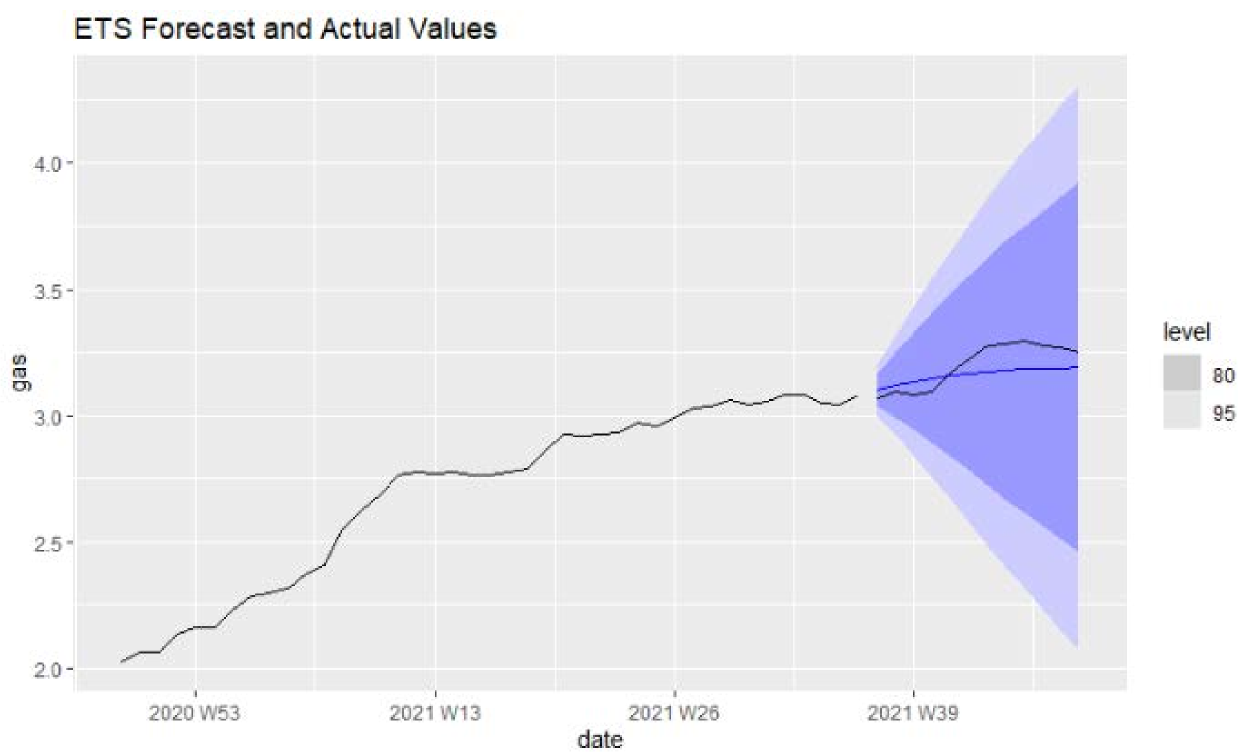

Model Comparison
4.0
3.5
3.0
2.5
2.0
gas
2020 W53
2021 W13
2021 W26
2021 W39
date
.model
ARIMA
dynreg
ETS
level
80
95

Fig 12


ETS Forecast and Actual Values
4.0
3.5
3.0
2.5
2.0
gas
2020 W53
2021 W13
2021 W26
2021 W39
date
level
80
95

Works Cited

Anderson, Soren T., Ryan Kellogg, James M. Sallee, and Richard T. Curtin. 2011. "Forecasting Gasoline Prices Using Consumer Surveys." *American Economic Review*, 101 (3): 110-14.

Andrea Bastianin, Marzio Galeotti, Matteo Manera, Forecasting the oil–gasoline price relationship: Do asymmetries help?, Energy Economics, Volume 46, Supplement 1, 2014, Pages S44-S56, ISSN 0140-9883, https://doi.org/10.1016/j.eneco.2014.08.006.

Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3.

Ervin Čeperić, Saša Žiković, Vladimir Čeperić, Short-term forecasting of natural gas prices using machine learning and feature selection algorithms, Energy, Volume 140, Part 1, 2017, Pages 893-900, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2017.09.026.

U.S. Energy Information Administration, Crude Oil Prices: West Texas Intermediate (WTI) - Cushing, Oklahoma [DCOILWTICO], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DCOILWTICO, September 30, 2021.

U.S. Energy Information Administration, US Regular Conventional Gas Price [GASREGCOVW], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/GASREGCOVW, September 30, 2021.